Minimum description length approach for unsupervised spectral unmixing of multiple interfering gas species

Julien Fade, Sidonie Lefebvre, Nicolas Cézard

Minimum Description Length approach for unsupervised spectral unmixing of multiple interfering gas species

Julien Fade,[1,2,*] **Sidonie Lefebvre,**[1] **and Nicolas Cézard**[1]

[1] *French Aerospace Lab (ONERA), Chemin de la Hunière, 91 761 Palaiseau, France*

[2] *Institut de Physique de Rennes UMR 6251, Université de Rennes 1, Campus de Beaulieu, 35 042 Rennes, France*

[*] *julien.fade@univ-rennes1.fr*

Abstract: We address an original statistical method for unsupervised identification and concentration estimation of spectrally interfering gas components of unknown nature and number. We show that such spectral unmixing can be efficiently achieved using information criteria derived from the Minimum Description Length (MDL) principle, outperforming standard information criteria such as AICc or BIC. In the context of spectroscopic applications, we also show that the most efficient MDL technique implemented shows good robustness to experimental artifacts.

OCIS codes: 070.4790 (Spectrum analysis); 300.0300 (Spectroscopy); 010.1030 (Absorption); 010.0280 (Remote sensing and sensors); 280.1120 (Air pollution monitoring)

References and links

1. P. Weibring, C. Abrahamsson, M. Sjholm, J. N. Smith, H. Edner, and S. Svanberg, "Multi-component chemical analysis of gas mixtures using a continuously tuneable lidar system," Appl. Phys. B **79**, 525–530 (2004).

2. J. R. Quagliano, P. O. Stoutland, R. R. Petrin, R. K. Sander, R. J. Romero, M. C. Whitehead, C. R. Quick, J. J. Tiee, and L. J. Jolin, "Quantitative chemical identification of four gases in remote infrared (9-11 μm) differential absorption lidar experiments," Appl. Opt. **36**, 1915–1927 (1997).
3. G. Wysocki, R. Lewicki, R. Curl, F. Tittel, L. Diehl, F. Capasso, M. Troccoli, G. Hofler, D. Bour, S. Corzine, R. Maulini, M. Giovannini, and J. Faist, "Widely tunable mode-hop free external cavity quantum cascade lasers for high resolution spectroscopy and chemical sensing," Appl. Phys. B: Lasers and Optics **92**, 305–311 (2008).
4. U. Platt and J. Stutz, *Differential Optical Absorption Spectroscopy* (Springer Berlin / Heidelberg, 2008).
5. R. A. Hashmonay, R. M. Varma, M. Modrak, R. H. Kagann, and P. D. Sullivan, "Simultaneous measurement of vaporous and aerosolized threats by active open path FTIR," Unclassified Technical Report ADA449529, Arcadis Geraghty and Miller Research Triangle Park NC (2004).
6. J. Kasparian, M. Rodriguez, G. Méjean, J. Yu, E. Salmon, H. Wille, R. Bourayou, S. Frey, Y. André, A. Mysyrowicz, R. Sauerbrey, J. Wolf, and L. Wöste, "White-Light Filaments for Atmospheric Analysis," Science **301**, 61–64 (2003).
7. D. M. Brown, K. Shi, Z. Liu, and C. R. Philbrick, "Long-path supercontinuum absorption spectroscopy for measurement of atmospheric constituents," Optics Express **16**, 8457–8471 (2008).
8. P. S. Edwards, A. M. Wyant, D. M. Brown, Z. Liu, and C. R. Philbrick, "Supercontinuum laser sensing of atmospheric constituents," in "Proc. SPIE Vol.7323," (2009), p. 73230S.
9. E. R. Warren, "Optimum detection of multiple vapor materials with frequency-agile lidar," Appl. Opt. **35**, 4180–4193 (1996).
10. S. Yin and W. Wang, "Novel algorithm for simultaneously detecting multiple vapor materials with multiple-wavelength differential absorption lidar," Chinese Opt. Lett. **4**, 360–363 (2006).
11. J. Fade and N. Cézard, "Supercontinuum lidar absorption spectroscopy for gas detection and concentration estimation," in "Proc. 25th International Laser and Remote-sensing Conference," (2010), pp. 798–801.
12. J. Rissanen, *Information and Complexity in Statistical Modeling* (Springer, New-York, 2007).
13. R. A. Stine, "Model selection using information theory and the MDL principle," Sociological Methods Research **33**, 230–260 (2004).
14. C. D. Giurcaneanu, "Stochastic complexity for the detection of periodically expressed genes," in "Proc. of IEEE International Workshop on Genomic Signal Processing and Statistics," (2007), pp. 1–4.
15. H. Chen, T. Kirubarajan, Y. Bar-Shalom, and K. R. Pattipati, "MDL approach for multiple low-observable track initiation," in "Proc. SPIE Vol. 4728," (2002), pp. 477–488.
16. M. Hansen and B. Yu, "Model selection and the principle of minimum description length," Journal of the Amer-

ican Statistical Association **96**, 746–774 (2001).

17. C. L. Mallows, "Some comments on c_p," Technometrics **15**, 661–675 (1973).
18. H. Akaike, "A new look at the statistical model identification," IEEE Trans. on Automatic Control **19**, 716–723 (1974).
19. G. Schwartz, "Estimating the dimension of a model," Annals of Statistics **9**, 461–464 (1978).
20. D. P. Foster and E. I. G., "The risk inflation criterion for multiple regression," Annals of Statistics **22**, 1947–1975 (1994).
21. J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, vol. 15 of *Series in Computer Science* (World Scientific, Singapore, 1989).
22. J. Rissanen, "Fisher information and stochastic complexity," IEEE Trans. Inform. Theory **42**, 48–54 (1996).
23. M. Duhant, W. Renard, G. Canat, F. Smektala, J. Troles, P. Bourdon, and C. Planchat, "Improving mid-infrared supercontinuum generation efficency by pumping a fluoride fiber directly into the anomalous regime at 1995nm," Accepted for publication in CLEO/Europe and EQEC 2011 Conference Digest, (2011), p. CD9_1.
24. A. Berrou, M. Raybaut, A. Godard, and M. Lefebvre, "High-resolution photoacoustic and direct absorption spectroscopy of main greenhouse gases by use of a pulsed entangled cavity doubly resonant OPO," Appl. Phys. B: Lasers and Optics **98**, 217–230 (2010).
25. R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society B **58**, 267–288 (1996).
26. E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," Annals of Statistics **37**, 2145–2177 (2009).

1. Introduction

Air pollution monitoring in the atmosphere has motivated the development of many active optical instruments based on absorption spectroscopy. Ideally, a single instrument should be able to detect and quantify numerous gas species. It is therefore appropriate to use an illumination source that can cover a large spectral range. Two kinds of sources can be used, which are: i) narrow-line lasers with broad tunability, and ii) instantaneous broadband sources. Both families have demonstrated high potential for measurement of multi-components gas mixtures in the atmosphere. Narrow-line tunable lasers have been used in multi-wavelength systems

like DIfferential Absorption Lidars (DIAL) [1, 2] and Tunable-Diode Laser Absorption Spectroscopy (TD-LAS) [3]. Instantaneous broadband sources have been used in various experiments schemes such as Differential Optical Absorption Spectroscopy (DOAS) [4], open-path active Fourier-Transform InfraRed (FTIR) spectroscopy [5], white-light filament-induced spectroscopy [6], and more recently, supercontinuum fiber laser spectroscopy [7, 8]. These various techniques share a common experimental design which is sketched in Fig.1.

All these techniques provide multi-spectral absorption data that can be processed by multivariate statistical analysis in order to characterize the gas mixture. When the number and nature of the chemicals are a priori known, efficient algorithms can be designed to estimate their concentrations [9, 10, 11]. However, in many practical cases, the number, nature, and concentration of gas components are all unknown. In such situations, the same algorithms are inclined to overfit signal noise by assigning non-zero concentrations to many gas species in the fixed list of expectable gases (all of them being estimated at the same time). This results in complex and often unrealistic gas diagnosis. To avoid this, it is necessary to design unsupervised methods enabling simultaneous gas selection and concentration estimation. In this paper, we use the powerful concept of Minimum Description Length (MDL) principle to tackle this problem. We illustrate the potential of the method for spectral unmixing of several chemicals in the mid-infrared range. This spectral range is of particular interest for air pollution monitoring, as many industrial and greenhouse gases exhibit strong absorption lines in this band.

In broad outline, the MDL principle is based on the idea that the best model describing the measured data must minimize the code length needed to describe the data and to encode the model itself [12]. Such a principle has already been applied in various domains, such as social sciences [13], biology [14] or radar signal processing [15] for instance. For the first time to the best of our knowledge, we show that this principle can be used for spectroscopic applications. More precisely, the approach presented in this paper allows unsupervised spectral unmixing

of gas mixtures to be simply operated, with detection performances that outperform standard information criteria.

This paper is organized as follows: in the next section, we present the physical situation considered, and the principle of the MDL-based spectral unmixing algorithm is detailed. In section 3, we present and analyze some simulation results, allowing us to quantitatively compare the performance of the MDL-based approaches with standard methods. We also analyze the robustness of the proposed method when experimental outliers occur in the measurement process. Finally, the conclusion and perspectives of the paper are given in Section 4.

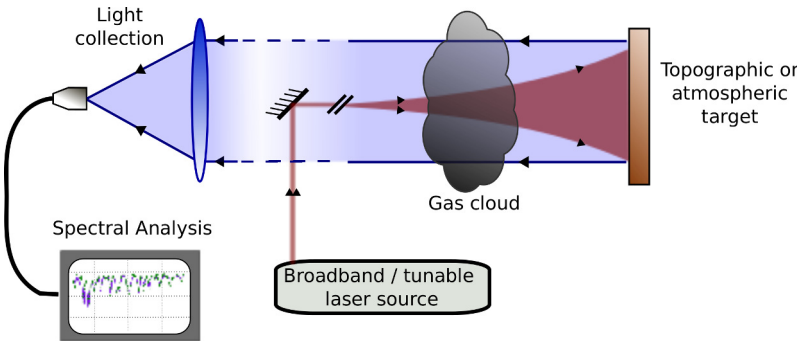

Fig. 1. Illustration of an absorption spectroscopy experiment using an active broadband illumination or tunable laser source.

2. Principle of unsupervised spectral unmixing algorithm

2.1. Posing of the problem

Before presenting the principle of the unsupervised spectral unmixing method addressed in this paper, let us detail the physical model that will be considered in the following. In most of absorption spectroscopy experiments, one is interested in measuring a vector $\mathbf{X}$ containing intensity measures on M spectral slits (or wavelengths) not necessarily adjacent. In the presence of absorbing gas species, these spectral measurements reveal specific absorption patterns depending on the nature and concentration of the chemicals encountered by the probe light beam.

These spectral absorption patterns are superimposed with the spectral baseline of the active illumination source. The vector $\mathbf{X}$ of the measured intensities is linked to the K-dimensional vector $\mathbf{c}$ containing the gases concentrations $\mathbf{c} = [\mathrm{c}_1, \ldots, \mathrm{c}_K]^T$ through the following equation

$$\mathbf{X} = \left(\mathbf{a}_{0\mathrm{u}} e^{-\mathbf{H_u c}}\right) * g, \tag{1}$$

where g denotes the spectral slit (or laser linewidth) convolution function, which is assumed known in the following. In this equation, $\mathbf{a}_{0\mathrm{u}}$ denotes the baseline spectrum, and the $M \times K$ matrix $\mathbf{H_u} = [\mathbf{h_u}_1, \mathbf{h}_2, \ldots, \mathbf{h_u}_K]$ contains the unconvolved high-resolution absorption spectra of the K gas species. For the sake of simplicity, we will only consider in this paper the case of small absorption optical depths (i.e., $\mathbf{H_u c} \ll 1$). Moreover, we assume that the baseline $\mathbf{a}_0$ is varying slowly with respect to both the absorption lines and the convolution function widths. In such conditions, the measured intensities can be written,

$$\mathbf{X} = \mathbf{a}_0 e^{-\mathbf{Hc}}, \tag{2}$$

where the matrix $\mathbf{H} = [(\mathbf{h_u}_1 * g), (\mathbf{h_u}_2 * g), \ldots, (\mathbf{h_u}_K * g)]$ contains the convolved absorption spectra of the K gas species, and where the convolved spectral baseline $\mathbf{a}_0$ is assumed known, either from instrumental calibration or with a precise radiometric model of the illumination source. More accurate models involving deconvolution procedures, as well as the influence of a possible resolution mismatch between the instrument and the model are outside the scope of this paper, but could deserve investigation in future work.

The noisy experimental intensity measures over the M spectral slits, obtained for instance with a dispersive spectrometer or a FTIR spectrometer, will be denoted $\tilde{\mathbf{X}}$ in the remaining of this paper. It is a common procedure to use the logarithm of the measured data so as to obtain a linear regression model of the following form:

$$\tilde{\mathbf{Y}} = \ln \tilde{\mathbf{X}} = \mathbf{b}_0 - \mathbf{H} \cdot \mathbf{c} + \mathbf{n}, \tag{3}$$

with $\mathbf{b}_0 = \ln \mathbf{a}_0$, and where the M-dimensional zero-mean random vector $\mathbf{n}$ allows us to model the experimental noise. We assume that the noise contribution to the measured signal $\tilde{\mathbf{Y}}$ can be correctly accounted for with a Gaussian additive model. We also assume independence between the noise affecting two distinct spectral slits, i.e., $\langle \mathrm{n}_i \, \mathrm{n}_j \rangle = 0$ if $i \neq j$. For such a linear regression model, the usual estimator is $\hat{\mathbf{c}} = (\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}^T(\tilde{\mathbf{Y}} - \mathbf{b}_0)$ and is usually referred to as the Minimum Mean Squared Error (MMSE) estimator since it minimizes the Residual Sum of Squares $RSS = (\tilde{\mathbf{Y}} - \hat{\mathbf{Y}})^T(\tilde{\mathbf{Y}} - \hat{\mathbf{Y}})$, with $\hat{\mathbf{Y}} = \mathbf{b}_0 - \mathbf{H}\hat{\mathbf{c}}$.

Under the hypothesis of Gaussian fluctuations and if the noise variance is not assumed *a priori* known during the estimation procedure, it can be shown [16] that this quantity is related to the loglikelihood $\ell_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{Y}}|\mathbf{H})$ of the observed data through the following equation

$$\ell_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{Y}}|\mathbf{H}) = \ln P_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{Y}}|\mathbf{H}) = -\frac{M}{2} \ln RSS + ct, \tag{4}$$

where ct denotes an additive constant independent of the measured data. It can be noted that this last equation shows that the MMSE is also the Maximum Likelihood (ML) estimator under Gaussian fluctuations.

2.2. Model selection

The issue of model selection arises in many practical situations. For the problem at hand, two questions have to be answered: how many gas components (regressors) do we need to describe the experimental data, and which regressors have to be selected in the linear regression model of Eq.(3) to best explain the observations ? Without any model selection step, the most exhaustive regression model would include any gas species presenting characteristic absorption lines within the spectral range considered, and may lead to misleading and imprecise (if not incorrect) results, mostly due to overfitting of the noise. To avoid such undesirable situations, many penalization methods have been proposed among which we can cite the Akaike Information Criterion (AIC) [18], the Bayesian Information Criterion (BIC) [19], the Risk Inflation Crite-

rion (RIC) [20], etc. These so-called *information criteria* make it possible to introduce sparsity constraints in the regression model, by selecting the solution (i.e., the regressor matrix $\mathbf{H}$) which minimizes $-\ell_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{Y}}|\mathbf{H})+\mathscr{C}$, with a different penalization term $\mathscr{C}$ depending on the information criterion considered. It can be noted however that since the loglikelihood is proportional to the logarithm of the RSS, up to an additive constant independent of the selected regression model [16], the model selection can be operated equivalently by minimizing $M/2\ln RSS+\mathscr{C}$.

Let us briefly recall two of the classical information criteria, which will be used in the remaining of the paper as benchmarks to assess the quality of the proposed MDL-based methods. The simplest is the Akaike Information Criterion (AIC) [18], which introduces a penalization term equal to the number K of regressors included in the model. In the case of samples of limited size, this penalization term can be refined and is usually referred to as AICc and will be denoted $\mathscr{C}^{(a)}$ in the following, with [16]

$$\mathscr{C}^{(a)}=\frac{M}{2}\frac{1+K/M}{1-(K+2)/M}. \tag{5}$$

We shall also consider the well-known Bayesian Information Criterion (BIC) [19], whose penalization term reads

$$\mathscr{C}^{(b)}=\frac{K}{2}\ln M. \tag{6}$$

Other information criteria can be found in abundance in the literature, which may suggest that an appropriate "most efficient" criterion at hand can be designed for a given statistical problem. Among various attempts to build a general theoretical framework to interpret model complexity, the Minimum Description Length (MDL) principle introduced by Rissanen [12] is an interesting and fruitful approach. The MDL principle is based on the underlying idea that the best description of the data will be given by the model leading to the shortest code length (expressed in bits or in *nats* (1 *nat*= $\ln 2$ bits)) needed to both describe the data given the model, and to encode the model itself [12, 16].

It is interesting to note that one of the simplest forms of the MDL, the so-called *two-stage description length*, is intimately related to the BIC. Indeed, assuming a M-dimensional data described with a probability density function (Pdf) parametrized with a K-dimensional vector θ, it can be shown that the code length (in *nats*) needed to describe the data is given by the negative loglikelihood ($-\ell(\tilde{\mathbf{Y}}|\theta)$) [12, 16]. In addition, the code length needed to describe such a model with K parameters can be shown to be equal to approximately $K/2\ln M$ [12, 16]. From this result, it can thus be seen that minimizing the two-stage MDL boils down to applying a BIC penalization.

More recently, sophisticated forms of the MDL principle have been proposed, with a constant effort towards loosening the assumptions held on the observed data. We shall focus in the following on two MDL approaches whose expressions are recalled below. Detailed theoretical foundings of these MDL theories can be found in Refs [12, 16, 21].

Mixture MDL and g-prior (gMDL): Within the framework of *mixture* MDL [21], a prior distribution is assigned to the vector parameter θ. With a specific choice of the prior distribution (Zellner's g-prior), one obtains the so-called *gMDL* for which the criterion to minimize has the following closed form expression [16]:

$$\min \begin{cases} \frac{M}{2}\ln RSS + \mathscr{C}^{(g)} & \text{if } F > 1 \\ \frac{M}{2}\ln(\tilde{\mathbf{Y}}-\mathbf{b}_0)^T(\tilde{\mathbf{Y}}-\mathbf{b}_0) & \text{otherwise,} \end{cases} \tag{7}$$

where $F = (M-K)\big[(\tilde{\mathbf{Y}}-\mathbf{b}_0)^T(\tilde{\mathbf{Y}}-\mathbf{b}_0) - RSS\big]/K\,RSS$ is the standard F-ratio for testing the null model containing the spectral baseline only. The penalization term $\mathscr{C}^{(g)}$ in Eq.(7) is given in [16] and can be written

$$\mathscr{C}^{(g)} = \frac{K}{2}\ln F + \frac{M}{2}\ln\frac{M}{M-K}. \tag{8}$$

Normalized Maximized Likelihood (nMDL): Lastly, we shall be interested in the recently proposed Normalized Maximized Likelihood form of the MDL [22]. This approach has proved

efficient in various practical problems and has shown several optimality properties [12, 16]. For the statistical problem considered in this paper, the nMDL theory suggests to introduce the following penalization terms [12]:

- For a model including the baseline only:

$$\mathscr{C}^{(n)} = \frac{M}{2}\ln\frac{2\pi}{M} + \frac{1}{2}\ln\frac{M}{2} + \ln\ln\frac{b}{a}. \tag{9}$$

- In any other cases:

$$\mathscr{C}^{(n)} = \frac{K}{2}\ln F + \frac{1}{2}\ln K(M-K) + \frac{K}{2}\ln M + \frac{M}{2}\ln\frac{2\pi}{M-K} + 2\ln\ln\frac{b}{a} - \ln 2 + \mathscr{L}_c, \tag{10}$$

  where $\mathscr{L}_c$ denotes the code length needed for encoding the model. Following Rissanen, we use the code length

$$\mathscr{L}_c = \min\left\{K_{max}, \left[\ln\frac{K!(K_{max}-K)!}{K_{max}!} + \ln K + \log_2\ln(e\,K_{max})\right]\right\} \tag{11}$$

  for a selection among K_{max} potential regressors contained in the spectral database.

It must be noted that the nMDL approach requires the hyperparameters a and b to be estimated. According to Rissanen's indications [12], the estimator of the hyperparameter a is given by the RSS obtained with the most exhaustive model (i.e. K_{max} regressors included) while the estimator of the hyperparameter b is the RSS obtained with the less exhaustive model (i.e. baseline only).

2.3. Stepwise algorithm for unsupervised spectral unmixing

Whatever the criterion selected, the determination of the optimal model requires an exhaustive search among all possible models which is computationally intensive if the number of potential regressors K_{max} is important. Instead of carrying out extensive operational research techniques such as *branch & bound* for instance, we implement a stepwise search algorithm for the sake

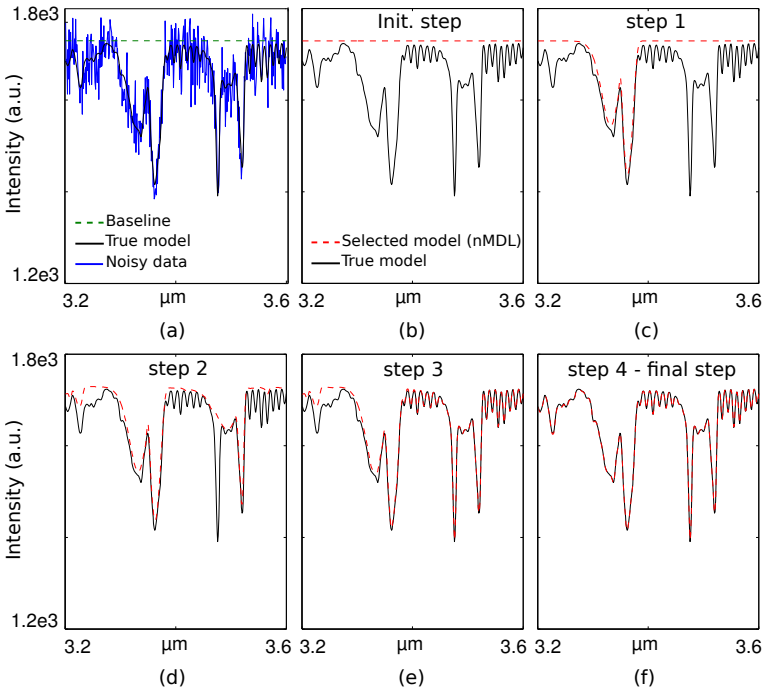

Fig. 2. (a) Example of simulated noisy data with S-SNR=6.3 dB (blue curve) superimposed with the true spectrum (black curve) and baseline (green dashed curve). (b)-(f) Comparison of the reconstructed signal after various steps of nMDL-based stepwise model selection (red dotted curve) with the true spectrum (black curve).

of computation rapidity. We use a forward stepwise algorithm with an initialization state containing the baseline only. At each step of the algorithm, the regressor (i.e., the gas species) that most diminishes the criterion is included in the model, until any further increase in the model complexity leads to an increase of the criterion. An example of iterative model selection is illustrated in Fig.2. At each step of the algorithm implementing nMDL criterion, the corresponding reconstructed signal is plotted in red dashed curve, and can be compared with the signal obtained with the true model (black curve).

Since we are concerned with absorption spectroscopy applications, we also implement a modified version of the algorithm so as to include a positivity constraint in the estimation results by rejecting models leading to physically unwanted negative concentration values.

It can be noted that this algorithm could be further refined for future developments by including backward elimination steps to reduce the risk of reaching local minima. Nevertheless, as will be shown in the next section, the algorithm implemented here is sufficient to compare the quality of the MDL approaches and standard information criteria such as AICc and BIC for unsupervised spectral unmixing of gas components.

3. Implementation and comparison of MDL-based information criteria

3.1. Simulated absorption spectroscopy experiment

We simulated a typical absorption spectroscopy experiment by numerically generating spectral measurements over $M = 400$ adjacent spectral slits, spanning between 3.2 and 3.6 μm, with a simulated instrumental spectral resolution of 2.3 nm (Gaussian slit function). The physical situation considered in this experiment consisted of a spectrally uniform illumination propagated through a gas mixture with 4 components: O_3 (6000 ppm.m), NO_2 (500 ppm.m), CH_4 (70 ppm.m) and H_2CO (30 ppm.m), where the numerical values in brackets correspond to their respective path-length integrated concentration.

The model selection was operated with the stepwise algorithm presented above from a spectral database containing $K_{max} = 16$ gas species, including the 4 gases of the "true" model and 12 spectrally interfering species (such as H_2O, N_2O, NH_3, HCl, etc.) with significant absorption strength within the spectral range considered. The strong spectral overlap of the database species can be checked in Fig.3, where the absorption spectra of 8 gas species (among 16 in the spectral database) are plotted. In this figure, the spectra are convolved with a Gaussian kernel to match the spectral resolution of the instrument considered in the simulated experimental data.

To account for experimental/detection noise, M statistically independent realizations of Gaussian random noise with variance σ^2 were added to the absorption spectra generated over M spectral slits. Varying the noise variance allowed us to simulate experiments with different

values of the Signal to Noise Ratio (SNR), usually defined in the context of additive Gaussian noise as the ratio of the flat baseline value to the noise standard deviation σ. However, this quantity is poorly adapted to assess the difficulty of the estimation problem considered, since it only depends on the active illumination power, and does not depend on the absorption strength of the gas mixture to be detected. We thus introduce another figure of merit, denoted S-SNR for *spectral* SNR, and defined as:

$$\text{S-SNR} = \frac{\sqrt{\frac{1}{M}(\mathbf{b}_0 - \mathbf{Y})^T(\mathbf{b}_0 - \mathbf{Y})}}{\sigma} = \frac{\sqrt{\frac{1}{M}\mathbf{c}^T\mathbf{H}^T\mathbf{H}\mathbf{c}}}{\sigma}. \tag{12}$$

In this expression, the numerator can be interpreted as the root mean square of the absorption signal $\mathbf{b}_0 - \mathbf{Y} = \mathbf{H}\mathbf{c}$ from which the nature and concentration of the gas components have to be estimated. An increase of the gas mixture concentrations accentuates the spectral absorption patterns in the measured spectrum, thus leading to an easier identification/estimation. In that case, it can be seen from the above definition that the S-SNR value is correspondingly increased. An example of simulated noisy data is given in Fig.2.a for a S-SNR=6.3 dB.

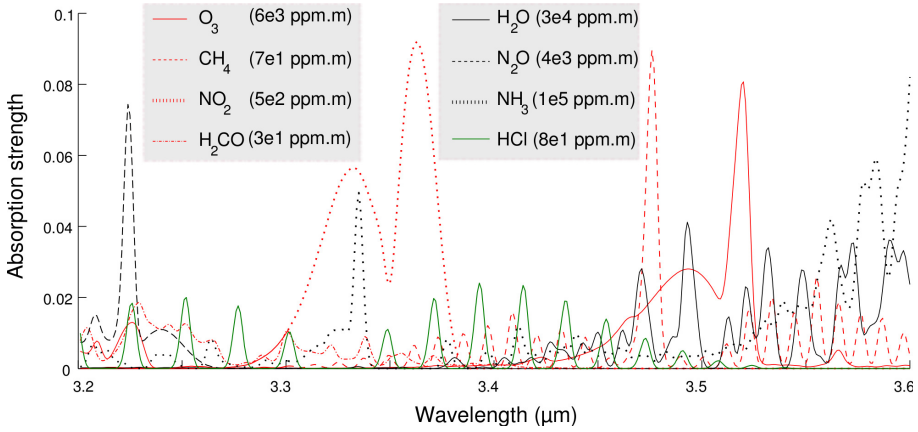

Fig. 3. Absorption spectra of the 4 gas components present in the mixture (red curves) and of 4 other chemicals of the spectral database (black and green curves) with resolution 2.3 nm. The green curve corresponds to the absorption spectrum of HCl, which is used in section 3.4 to simulate anomalous measurements (outliers).

3.2. *Simulation results*

The results of the numerical simulations are summarized in Table 1, where the percentage of correct model selections is given for the 4 information criteria compared in this paper and for different SNRs. For each physical situation considered, this percentage is evaluated over $R = 5.10^3$ realizations of the selection/estimation task on statistically independent simulated data. Two situations were considered according to whether light has undergone absorption from the gas mixture or not.

This table clearly reveals that in the context of unsupervised spectral unmixing, the MDL approaches implemented outperform the classical information criteria such as AICc or BIC, for reasonably high values of the SNR. This general result can be refined by observing that when the gas mixture is present, the nMDL is by far the most efficient criterion, with less than 2% erroneous selected models when S-SNR $\geq$ 6.3 dB, while the standard BIC selects approximately 17% erroneous models in the same conditions and AICc is strongly ineffective, leading to a large majority of erroneous selections. For lower values of the signal to noise ratio (S-SNR $<$4.3 dB) however, the percentage of correct models selected by nMDL strongly diminishes, and better performance is obtained with BIC. As for the gMDL approach, it can be noted that this criterion outperforms BIC for high SNRs (S-SNR$\geq$ 9.8 dB), but the advantage quickly drops out as the SNR decreases.

To complement this analysis, it is interesting to focus on the distribution of the size of the selected models. In Fig. 4.a, the histogram of the selected models sizes is plotted for the 4 criteria and for a S-SNR=6.3 dB. This figure reveals a clear tendency for AICc to overfit the noise patterns, thus leading to strongly overestimated model sizes. If the size distributions for BIC and gMDL are very similar, with approximately 16% of overestimated models ($K = 5$), it is however interesting to note that nMDL appears very efficient at avoiding overfitting, with only 1% of overestimated selections and 0.4% of selections with only $K = 3$ components.

Table 1. Percentage of correct models selected by the stepwise algorithm with four information criteria (AICc, BIC, gMDL, nMDL) and for various values of the SNR.

	With gas mixture				No gas mixture			
S-SNR	AICc	BIC	gMDL	nMDL	AICc	BIC	gMDL	nMDL
20.3 dB	16.5	83.6	96.0	>99.9	11.1	81.5	98.8	93.6
14.3 dB	18.6	83.7	92.8	99.9	11.0	80.3	99.2	93.1
9.8 dB	18.5	83.8	87.6	99.8	10.3	81.7	99.0	93.6
6.3 dB	17.7	83.1	80.8	98.9	10.6	82.0	99.0	92.3
4.3 dB	18.0	82.5	76.6	90.0	10.0	81.8	99.0	94.0
2.0 dB	16.4	74.0	63.5	53.5	10.8	81.7	98.9	92.8

This property has already been addressed in Ref.[12] and remains valid in the less favorable situations of low SNRs where nMDL is outperformed by BIC: when S-SNR=2 dB, nMDL leads to only 53.5% of correct models but more than 99% of the remaining selections have an underestimated size ($K = 3$) and the "missing" gas component is always H_2CO. In the context of absorption spectroscopy, this behavior seems interesting since it decreases the probability of erroneously detecting a gas component in excess and thus strengthens the confidence in the components selected with nMDL.

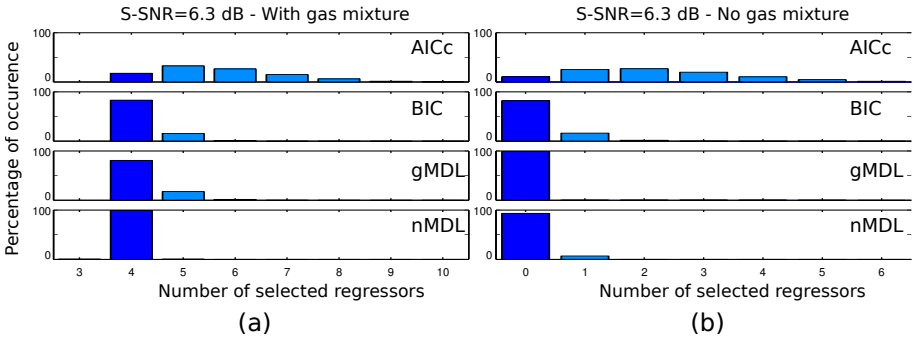

Fig. 4. Histograms of the number of regressors selected by AICc, BIC, gMDL and nMDL criteria for S-SNR=6.3 dB, (a): with a 4-components gas mixture; (b): without gas mixture.

Let us now analyze the second physical situation considered in the simulations where the illumination beam does not undergo any absorption. In this situation, it appears clearly again that MDL approaches lead to better results, when compared with standard criteria such as AICc and BIC. Once again, this result can be interpreted from the ability of MDL approaches to avoid overfitting, which can also be checked on the histograms of Fig.4.b. With approximately 99% of correct models, the gMDL criterion leads to the lowest probability of false alarm Pfa $\simeq 1 - 0.99 = 1\%$, which we define as the probability of detecting any gas mixture when there is not. On this particular point, the nMDL criterion appears less efficient with a Pfa $\simeq 6.5\%$.

From the above results, an interesting strategy for a practical implementation could be to use the *gMDL* criterion to test the null hypothesis. In case this hypothesis is rejected, the algorithm could switch to the nMDL criterion, which showed the best performance for model selection. In the next subsection, we analyze how a positivity-constrained implementation of the stepwise algorithm influences the previous results.

3.3. Influence of a positivity constraint

As stated in section 2.3, we also implemented a positivity-constrained version of the stepwise algorithm to provide physically acceptable results in the context of absorption spectroscopy. As can be checked in Table 2, such a constraint noticeably improves the quality of model selection with all the criteria considered. For instance, with a S-SNR=6.3 dB, the positivity-constrained algorithm selects 42.4% of correct models with AICc, 91.2% with BIC, 90.2% with gMDL and 99.3% with nMDL. When there is no gas mixture, the proportion of erroneous rejection of the null model hypothesis also diminishes whatever the criterion considered.

It can be noticed however that the performance of nMDL is less influenced by this constraint than the other criteria. This property may indicate that the nMDL criterion is intrinsically efficient at avoiding non-physical results in the context addressed here, even if no positivity-

Table 2. Percentage of correct models selected by the stepwise algorithm implementing a positivity constraint on the regression coefficients (i.e., on the gas components concentrations).

	With gas mixture				No gas mixture			
S-SNR	AICc	BIC	gMDL	nMDL	AICc	BIC	gMDL	nMDL
20.3 dB	40.0	92.6	98.4	>99.9	31.2	89.2	99.4	96.9
6.3 dB	42.4	91.2	90.2	99.3	28.6	89.3	99.50	97.0
4.3 dB	40.8	90.4	87.0	90.6	29.2	88.0	99.4	97.1
2.0 dB	41.6	81.8	77.6	57.3	29.1	88.8	98.4	96.9

constraint is applied to the algorithm.

3.4. Influence of outliers

To complete our analysis, we study the influence of measurement outliers. In practical situations of in-field experiments, many sources of measurement artifacts may exist, and it is likely that some amount of anomalous measures may occur. It is thus interesting to check the robustness of the implemented methods to the occurrence of outliers.

For that purpose, simulations were carried out in the same physical conditions as in the previous section, but the simulated data were generated in that case from the averaging of $N = 20$ independent measures. Among these $N = 20$ measures, we included a varying proportion of outliers, corresponding to the simulated noisy absorption spectrum of a single interferent gas species (HCl [80 ppm.m]), whose absorption spectrum is represented in green curve in Fig.3.

The results obtained are summarized in Table 3, where the percentage of correct models is given for a S-SNR=6.3 dB on the averaged signal. Once again, it can be clearly seen that in the context of spectral unmixing of interfering gas species, the nMDL criterion outperforms the other methods, with still 90% of correct models for a significant amount of outliers (20%). It

Table 3. Percentage of correct models selected by the stepwise algorithm with a S-SNR=6.3 dB for a varying proportion of measurement outliers.

	With gas mixture				No gas mixture			
% outliers	AICc	BIC	gMDL	nMDL	AICc	BIC	gMDL	nMDL
0 %	17.7	83.1	80.8	98.9	10.6	82.0	99.0	92.3
5%	13.7	80.0	75.7	97.4	8.3	77.1	98.1	93.8
20 %	7.3	62.2	53.6	90.1	5.2	59.7	92.4	93.9

can also be noted that the inclusion of outliers does not influence the Pfa obtained with nMDL (approximately $1 - 0.933 \simeq 6,7\%$) while this quantity noticeably decreases when other criteria are implemented.

4. Conclusion

In this paper, we presented an original technique for unsupervised spectral unmixing of multiple gas species. More precisely, we have shown that two Minimum Description Length approaches can be successfully implemented in a stepwise model selection algorithm. Applied on spectroscopic data, this algorithm allows one to estimate the number, nature and concentration of the components of an unknown gas mixture without requiring ajustment of any parameter.

In the context addressed in this paper, numerical simulations have demonstrated that the MDL approaches outperform the standard information criteria tested (AICc, BIC). When a gas mixture is present within the path of the illumination beam, the gMDL approach does not provide great improvement in comparison to classical BIC, but we illustrated its efficiency in avoiding false alarms when no gas mixture is present. However, the most promising results for a practical implementation were obtained with the Normalized Maximized Likelihood (nMDL) approach, which seems to be a very interesting alternative to standard criteria, and can still be implemented with a simple algorithm. The nMDL criterion strongly outperforms the other

methods for reasonable values of the SNR and provides the best robustness to measurements artifacts.

A promising perspective to this work is the opportunity to apply this method to experimental spectroscopic data due to recent development in our laboratories of appropriate mid-infrared powerful sources with broadband spectrum [23], or with highly tunable operating wavelength [24]. It must be noted that this approach is not limited to the case of absorption spectroscopy, and could be also applied in many situations requiring spectral unmixing (Raman spectroscopy, hyperspectral data processing, etc.). A further analysis of the influence of the spectral resolution and of the noise model would be also a useful theoretical continuation of this work, as well as the study of detection performances. A comparison of the MDL-based model selection techniques presented in this paper with other parsimonious model selection methods such as the *lasso* approaches [25, 26] is another interesting perspective.